



COMP9321:

Data services engineering

Week 8: Data Analytics

Semester 2, 2018

By Mortada Al-Banna, CSE UNSW

Data Driven Organizations

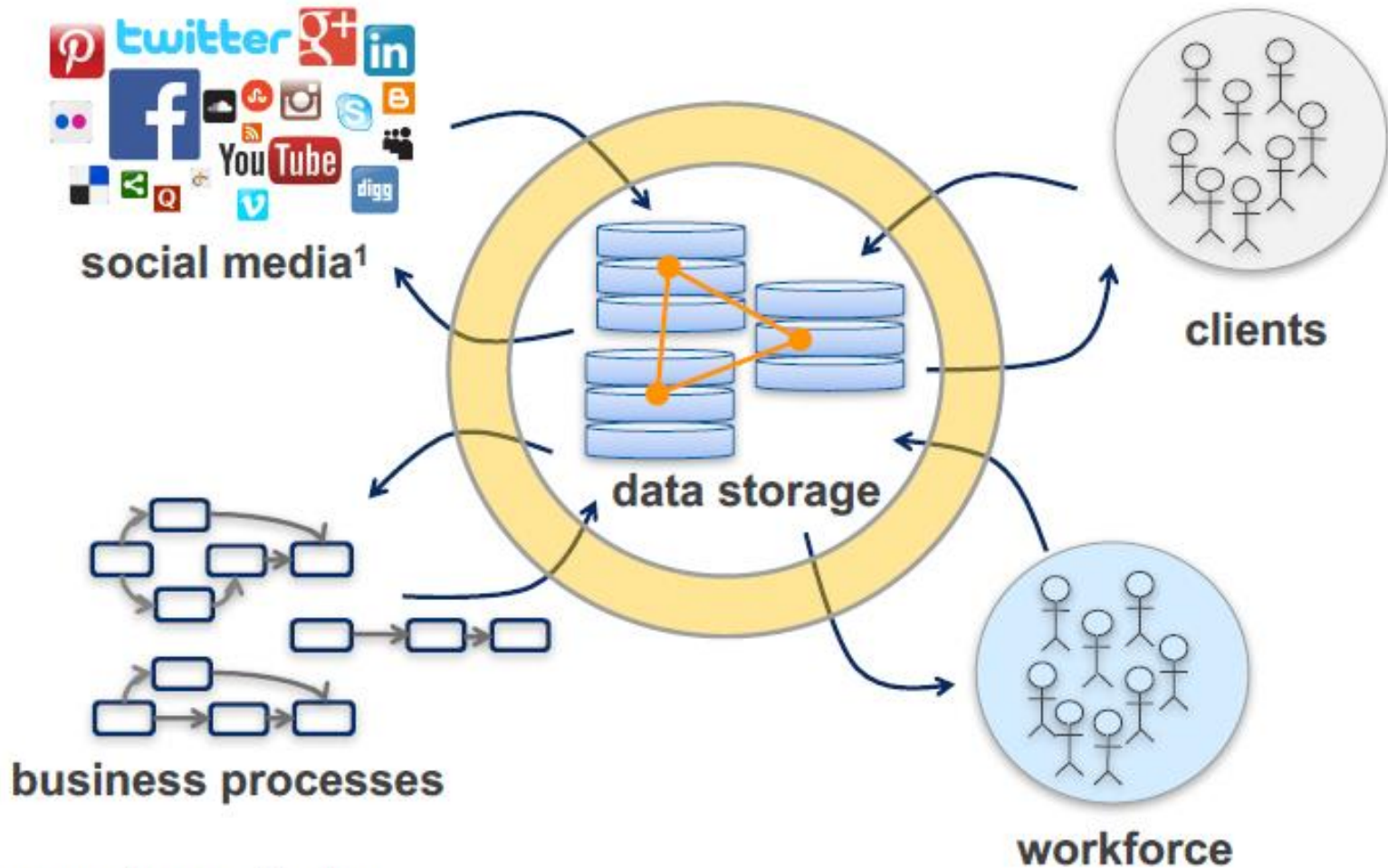


Image source: ¹commons.wikimedia.org

Data Driven Organizations and Data Analytics

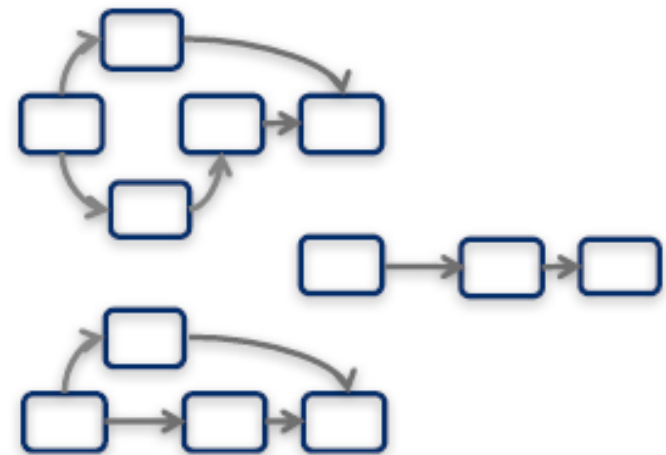
- Product and service recommendation
- Customer support
- Dashboard and reporting services
- Customer engagement
- Promotions and deals
- Product and service customization
- Communication



Clients

Data Driven Organizations and Data Analytics

- Key process performance indicators
- Process execution predictions
- Decision making support services
- Process mining
- Dynamic process adaptation
- People to task assignment
- Compliance verification



business processes

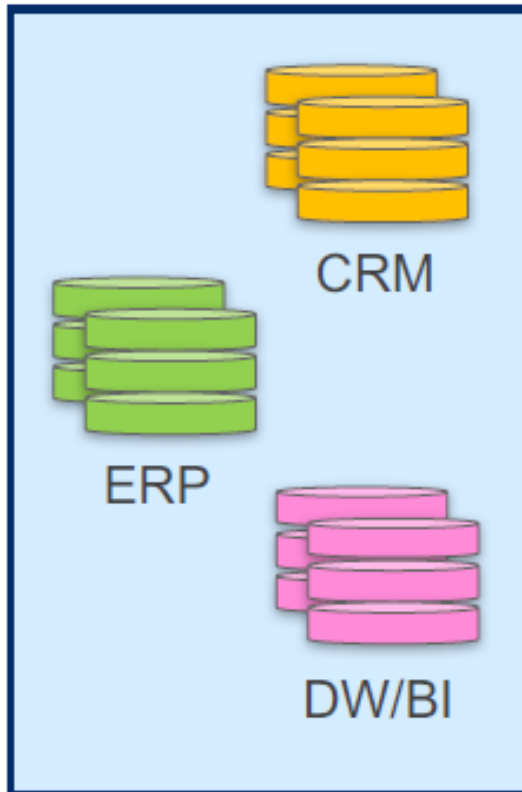
Data Driven Organizations and Data Analytics

- Product and service advertisement
- Sentiment analysis
- Demographics analysis
- Virality
- Social network insights

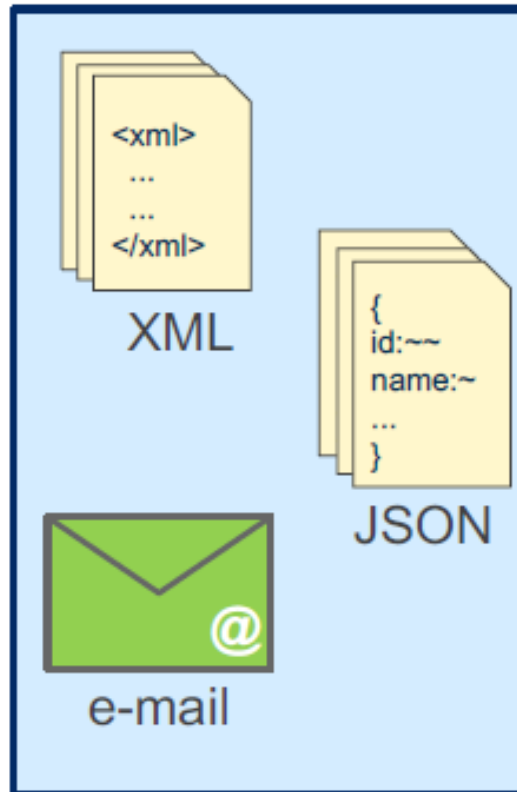


social media¹

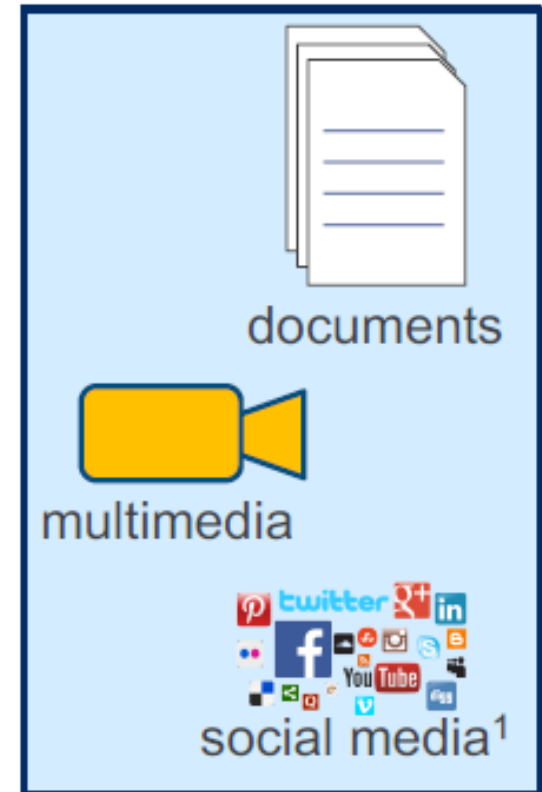
Data Used for Analytics



structured data



semi-structured data



unstructured data

Image source: ¹commons.wikimedia.org

Data Used for Analytics

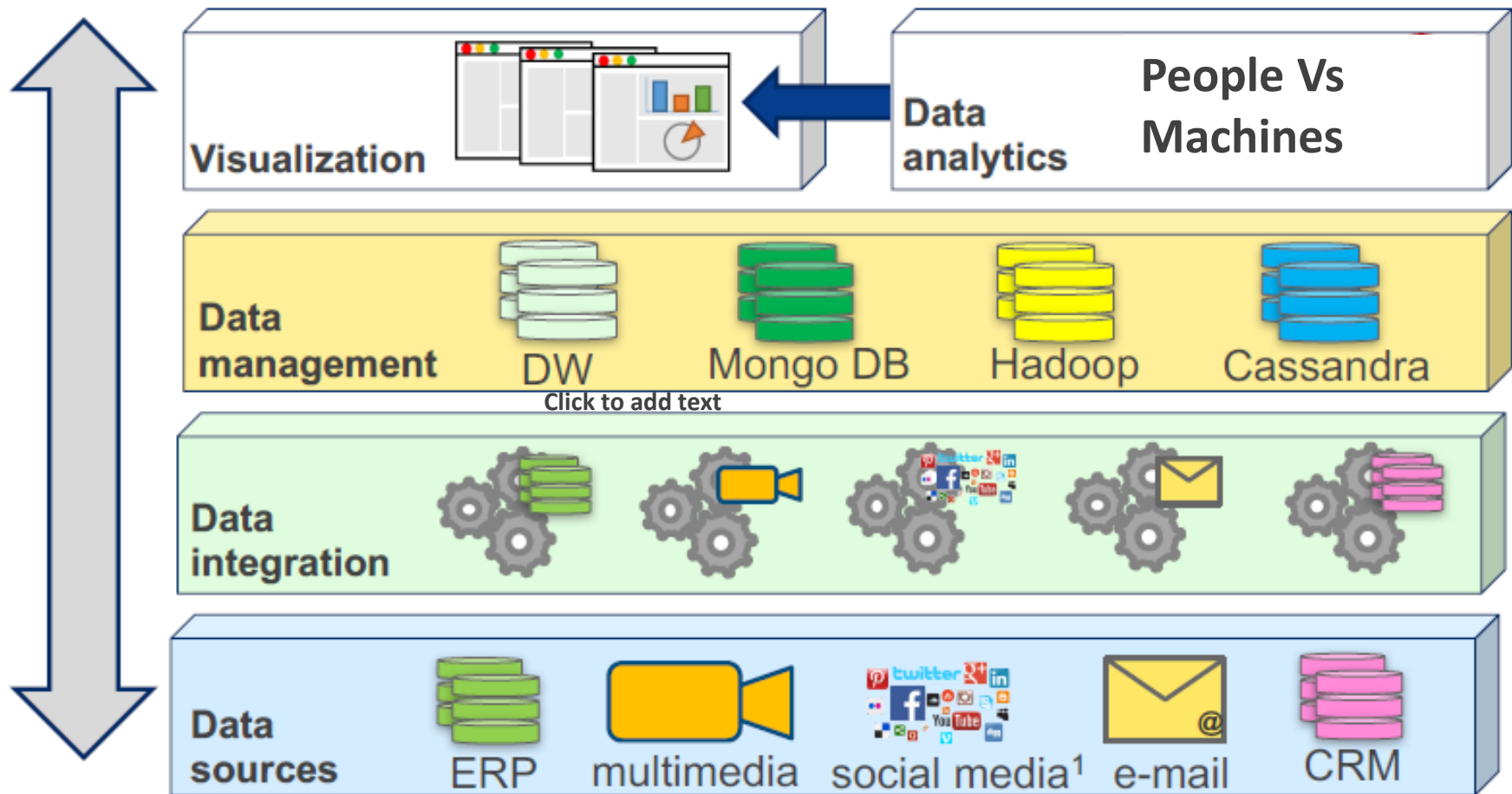


Image source: ¹commons.wikimedia.org

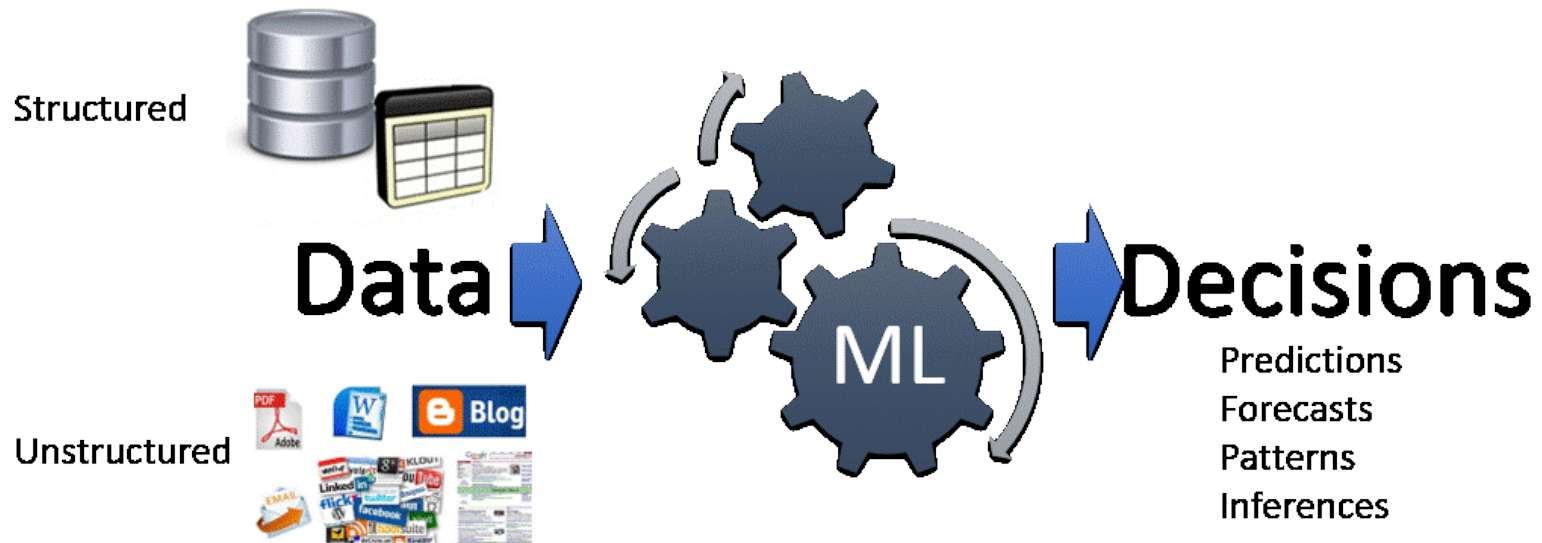
What is Machine Learning?

- Machine learning is an application of **artificial intelligence (AI)** that provides systems the ability to **automatically learn** and improve from experience without being explicitly programmed.
- Machine learning focuses on the development of “**computer programs that can access data and use it learn for themselves**”.

Useful Terminology

- Features
 - The number of features or distinct traits that can be used to describe each item in a quantitative manner.
- Samples
 - A sample is an item to process (e.g. classify). It can be a document, a picture, a sound, a video, a row in database or CSV file, or whatever you can describe with a fixed set of quantitative traits.
- Feature vector
 - is an n -dimensional vector of numerical features that represent some object.
- Feature extraction
 - Preparation of feature vector
 - transforms the data in the high-dimensional space to a space of fewer dimensions.
- Training/Evolution set
 - Set of data to discover potentially predictive relationships.

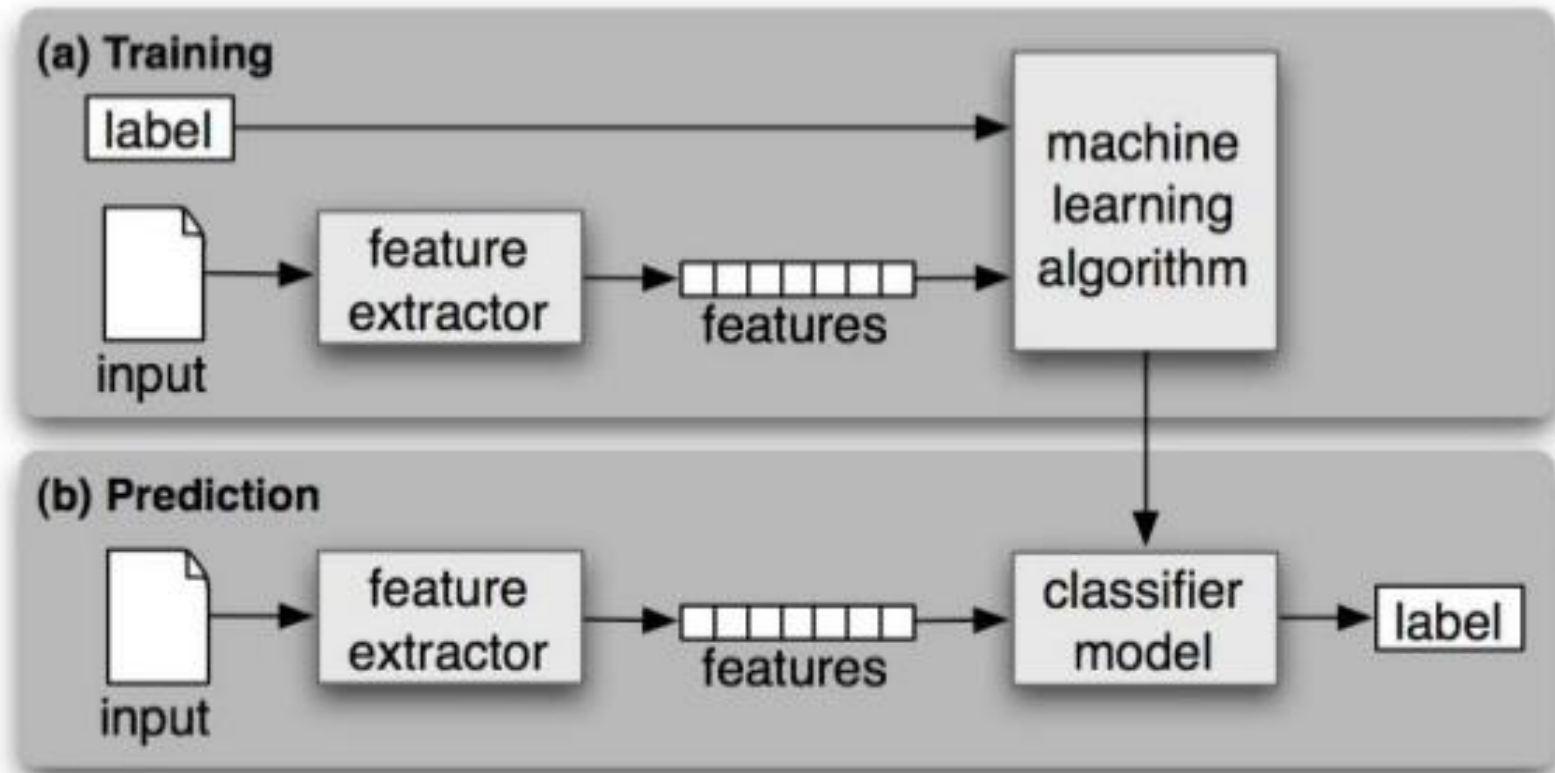
Machine Learning for Data Analytics



Machine Learning for Data Analytics

1. **Define** and **Initialize** a Model
2. **Train** your Model (using your training dataset)
3. **Validate** the Model (by prediction using your test dataset)
4. Use it: **Explore** or **Deploy** as a web service
5. **Update** and **Revalidate**

Example of a General Flow



What is an Apple?



Features:

1. Color: **Radish/Red**
 2. Type : **Fruit**
 3. Shape
- etc...



Features:

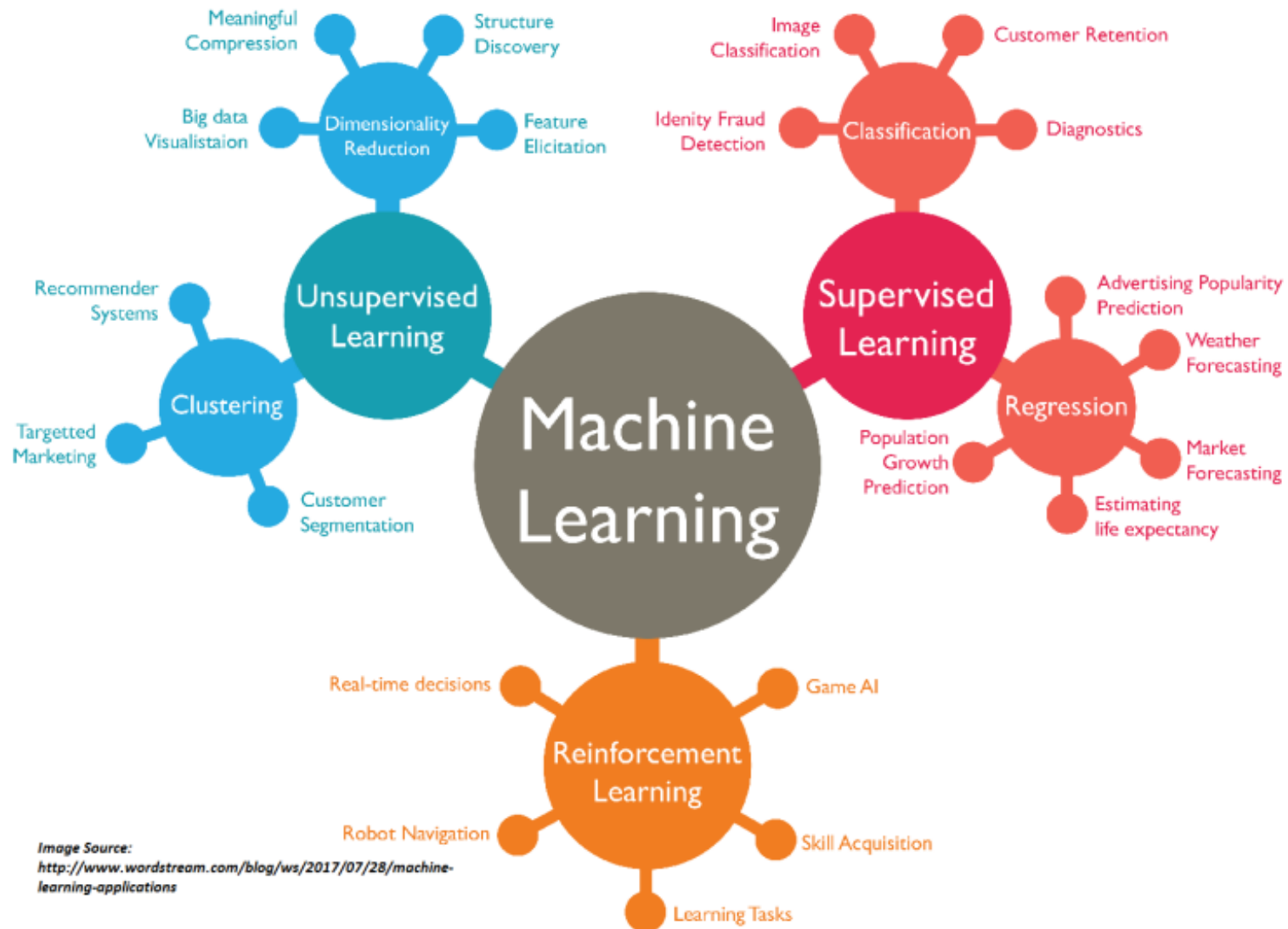
1. Sky Blue
 2. **Logo**
 3. Shape
- etc...



Features:

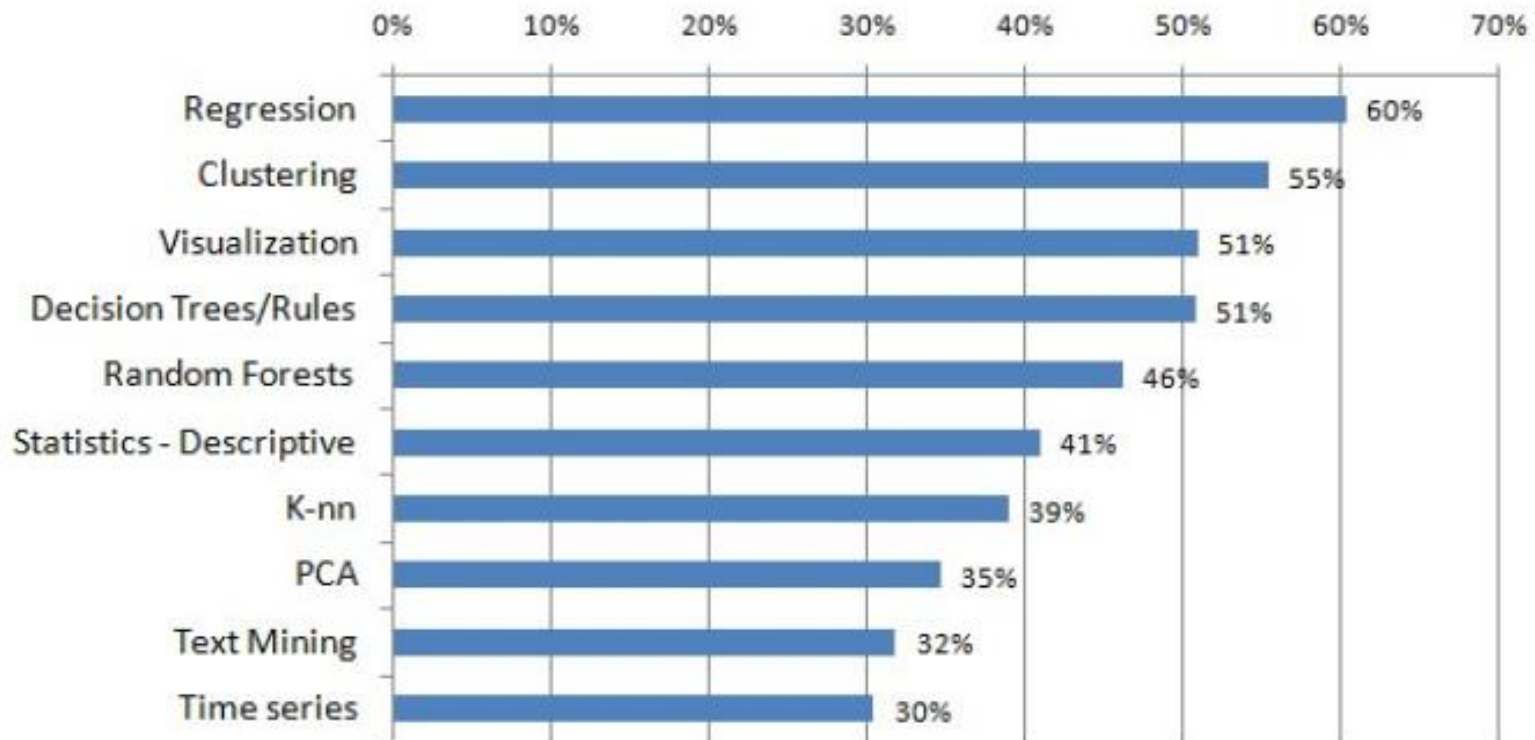
1. **Yellow**
 2. **Fruit**
 3. Shape
- etc...

Machine Learning Methods



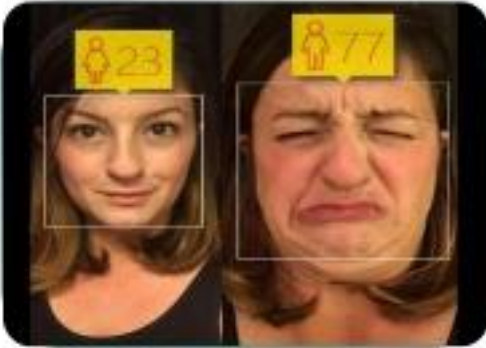
Machine Learning Methods

**Top 10 Data Science, Machine Learning Methods
Used, 2017**



Questions Machine Learning Can Answer

1. Is this A or B?



Classification Algorithms

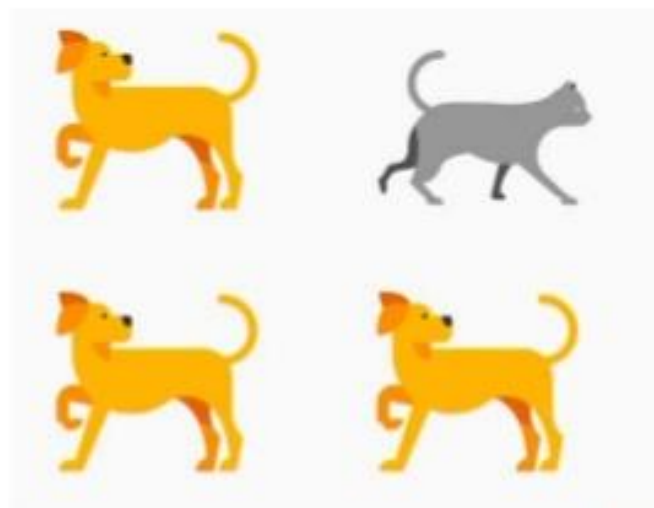


Questions Machine Learning Can Answer

2. Is this Weird?

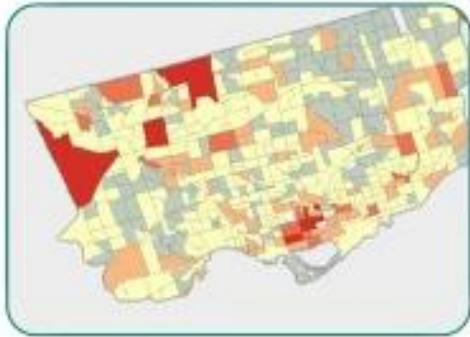


Anomaly detection algorithms



Questions Machine Learning Can Answer

3. How much? How many?



Regression algorithms



Questions Machine Learning Can Answer

4. How is this organized?



Clustering algorithms



Questions Machine Learning Can Answer

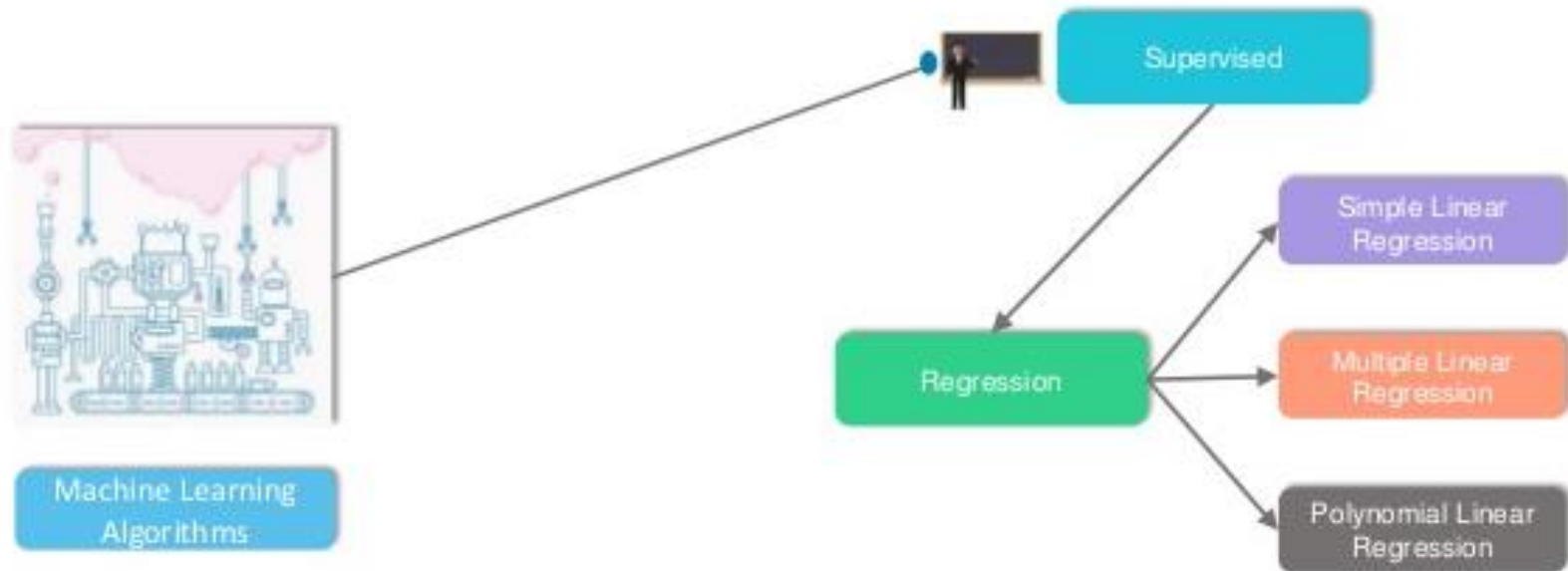
5. What should I do now?



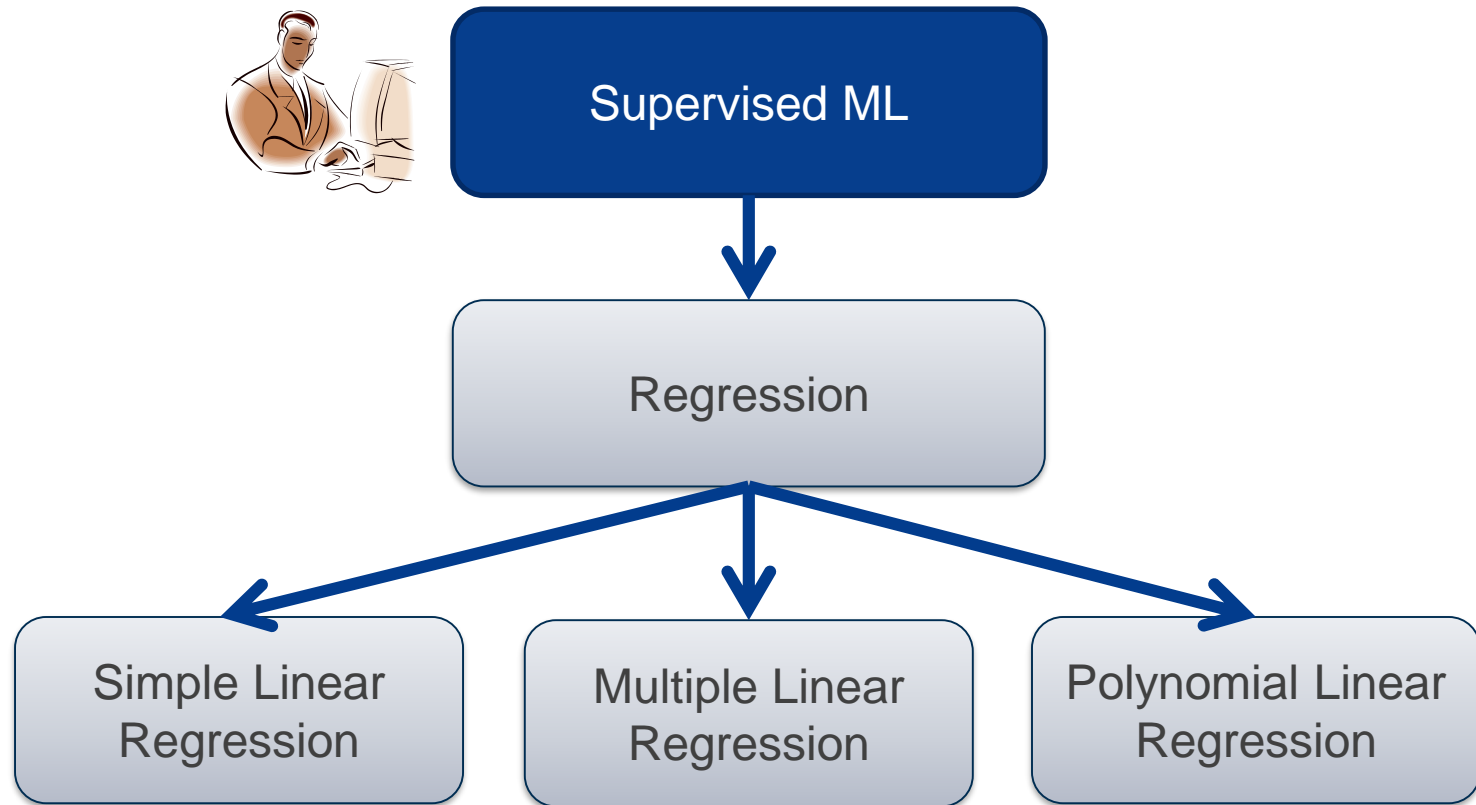
Reinforcement learning algorithms



Regression Analysis



Regression Analysis



Linear Regression (terminology)

- **Independent Variables (features):** An independent variable is a variable that is manipulated to determine the value of a dependent variable. Simply, they are the features which we want to use to predict some given value of Y . It can be also called an explanatory variable
- **Dependent Variable(target):** The dependent variable depends on the values of the independent variable. Simply put, it is the feature which we are trying to predict. This can also be commonly known as a response variable.

How Linear Regression Works

$$\hat{Y} = f(X) + \epsilon$$

X (input) = Assignment Results

Y (output) = Final Exam Mark

f = function which describes the relationship between X and Y

e (epsilon) = Random error term (positive or negative) with a mean zero (there are more assumptions for our residuals, however we won't be covering them)

Linear Regression Example

Training Set

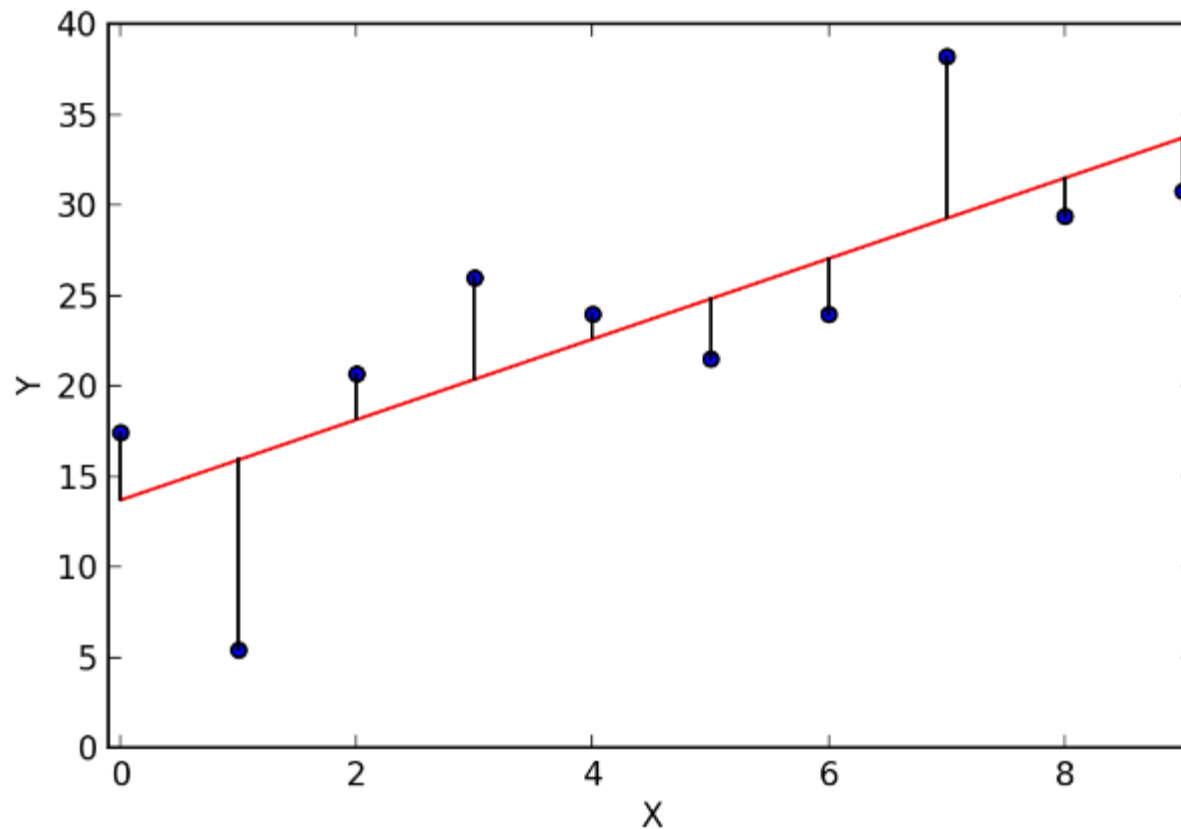
StudentID	Assignment_Mark (X)	Final_Exam_Mark (Y)
1292393	80	90
1823812	70	53
281823	63	74
....
183823	58	63
238381	54	61

Linear Regression Example

Test Set

StudentID	Assignment_Mark (X)	Final_Exam_Mark (Y)
184712	80	???
937217	70	???
...	...	???
836162	63	???

Linear Regression Example

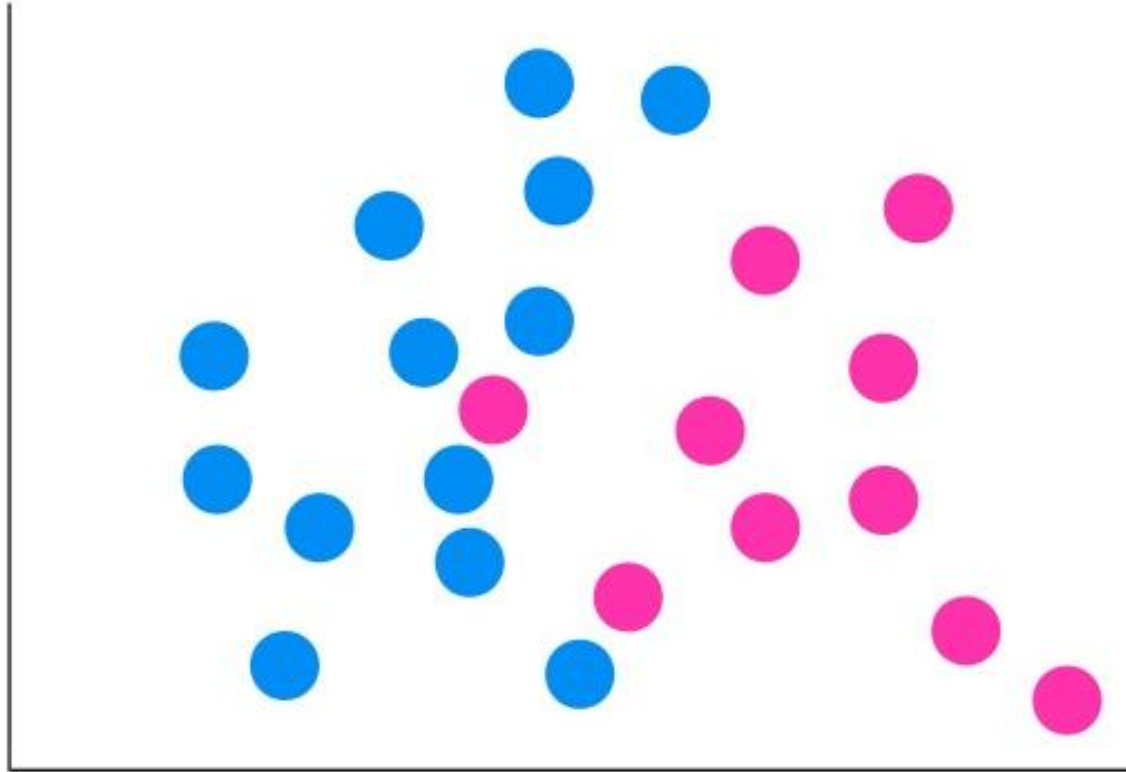


Where Y is our Final Exam Mark, and X is our Assignment Mark

Classification

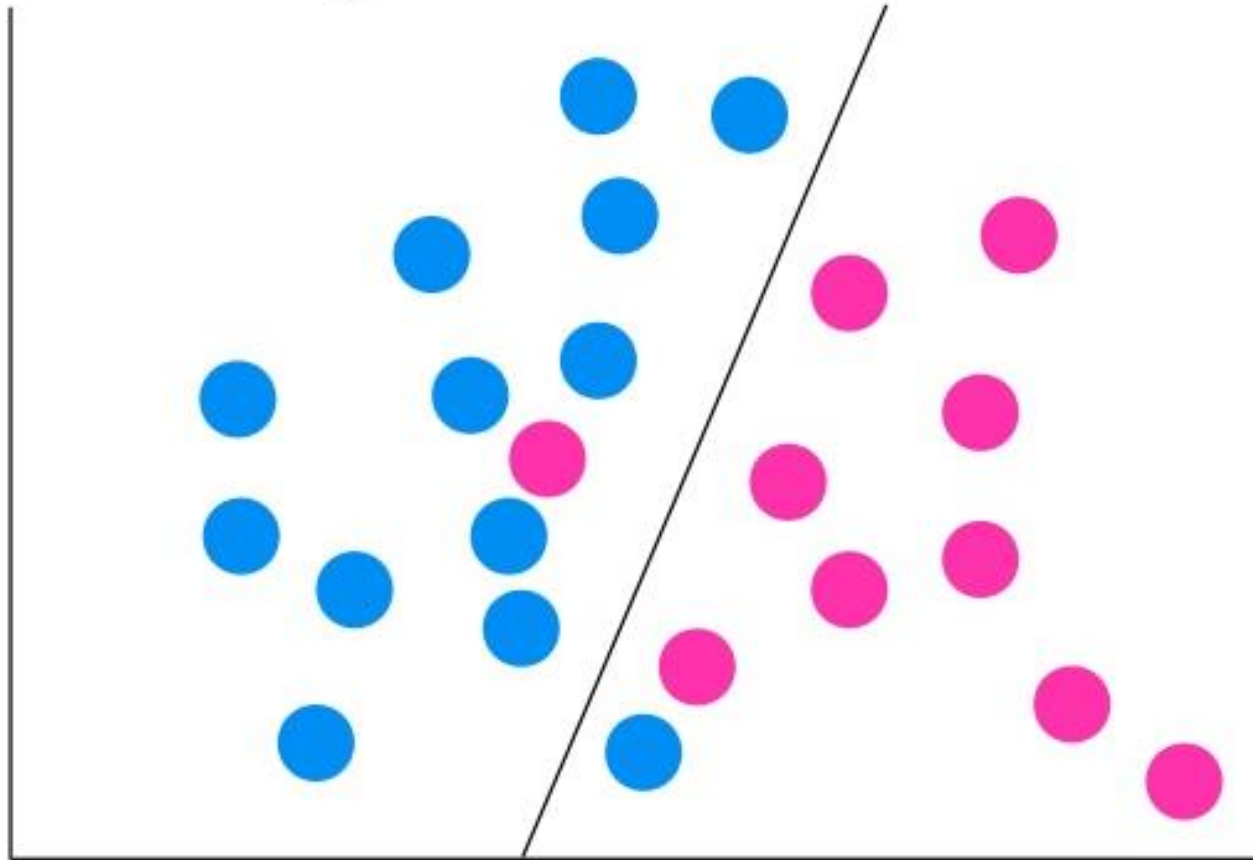
- Supervised Learning
- You need the data labelled with the correct answer to train the algorithm
- Trained classifiers then can map input data to a category.

Classification



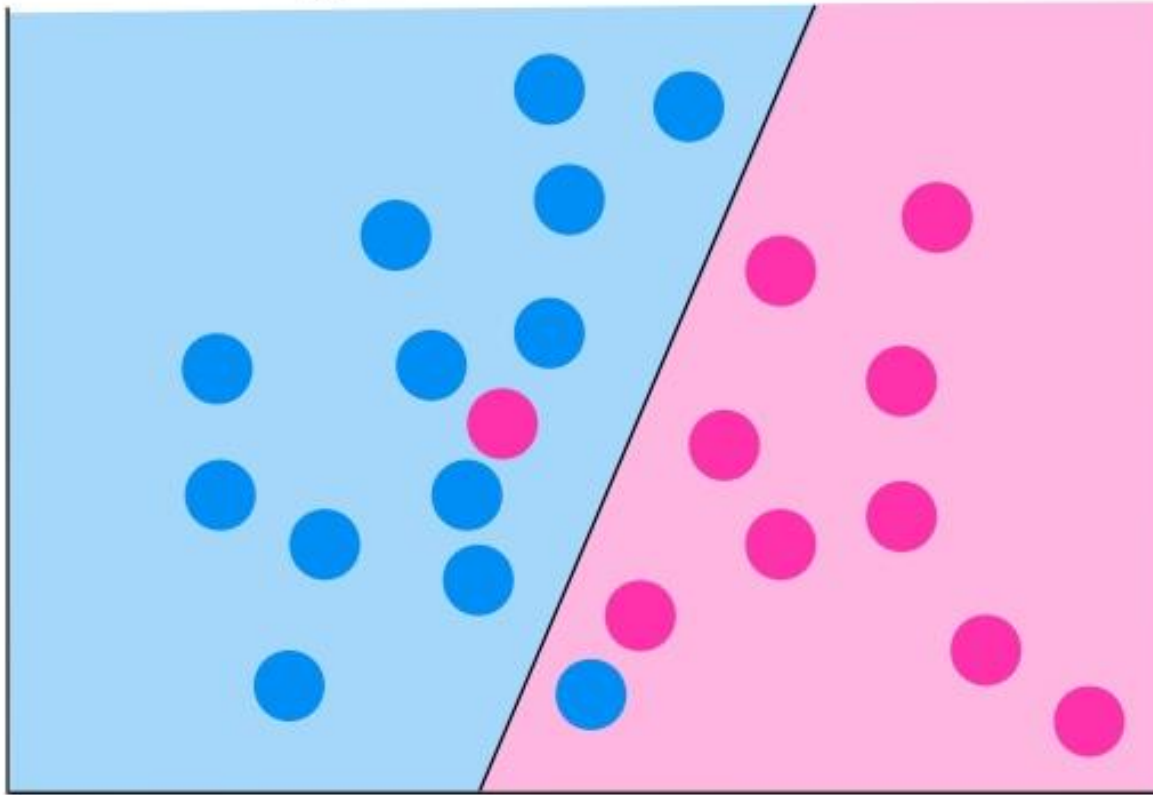
Classification

"draw a line through it"



Classification

"draw a line through it"



Linear Discriminants...

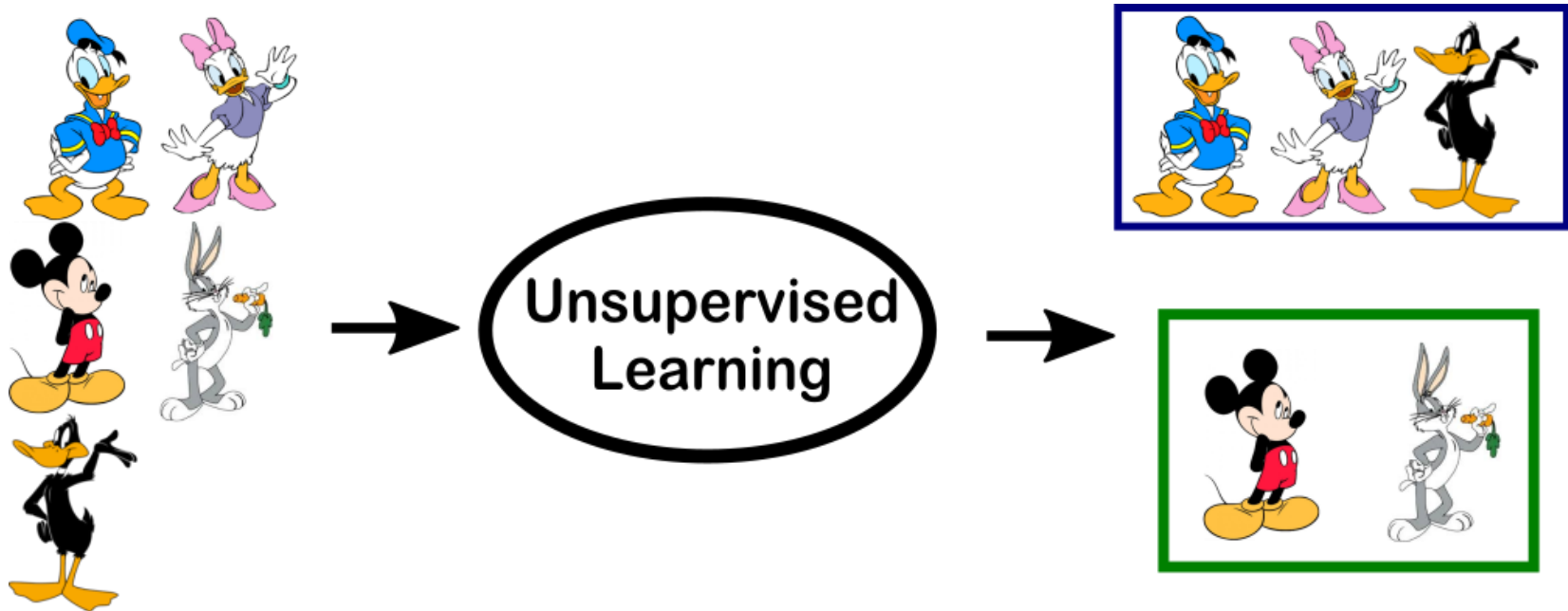
Probably Don't use these for all cases

We will introduce other options later

Clustering

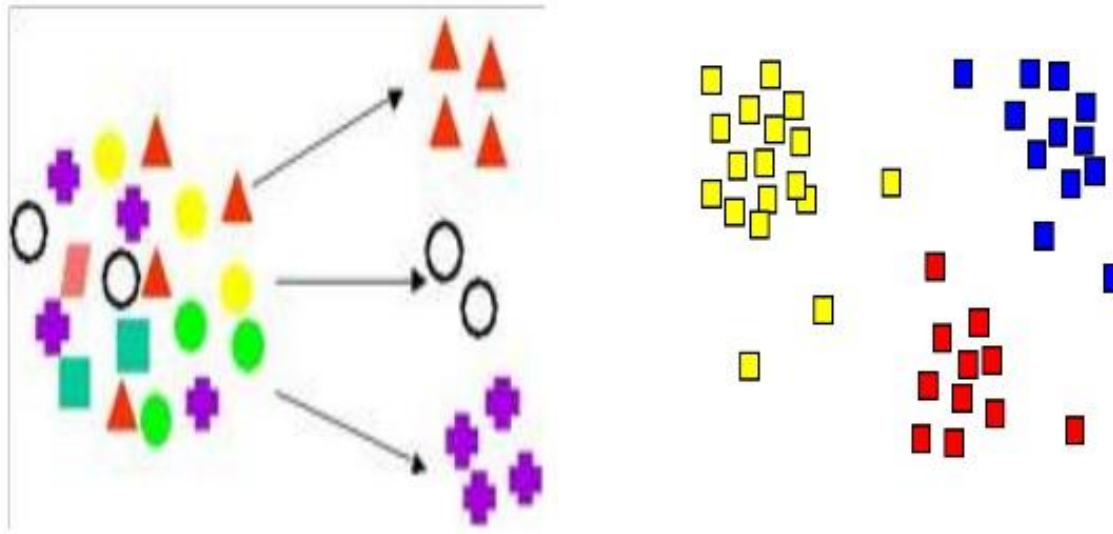
- Unsupervised Learning
- Automated grouping of objects into so called clusters
- Objects of the same group are similar
- Different groups are dissimilar

Clustering



Clustering

Examples of Clustering



Useful Tools



TensorFlow

Machine Learning Tools

Favorites

★
78

Stacks

495

I Use This

Fans	Votes	Jobs
406	52	246



scikit-learn

Machine Learning Tools

Favorites

★
13

Stacks

209

I Use This

Fans	Votes	Jobs
153	18	147



PredictionIO

Machine Learning Tools

Favorites

★
13

Stacks

31

I Use This

Fans	Votes	Jobs
35	4	0

Hacker News, Reddit,
Stack Overflow Stats

Y	Reddit	Stack Overflow
3.89K	3.26K	32.7K

Y	Reddit	Stack Overflow
-	912	12.3K

Y	Reddit	Stack Overflow
422	114	181

GitHub Stats

No public GitHub repository stats available

GitHub	★	🔗	📅
30.5K	15K	about 3 hours ago	

GitHub	★	🔗	📅
11.4K	1.87K	about 18 hours ago	



UNSW
SYDNEY

Useful Tools

- [TensorFlow](#)
- [scikit-learn](#)
- [PredictionIO](#)

Further Reading and Useful Resources

- <https://jakevdp.github.io/PythonDataScienceHandbook/05.02-introducing-scikit-learn.html>
- <https://towardsdatascience.com/introduction-to-machine-learning-db7c668822c4>
- <https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning>
- <http://gael-varoquaux.info/scikit-learn-tutorial/>

Q&A